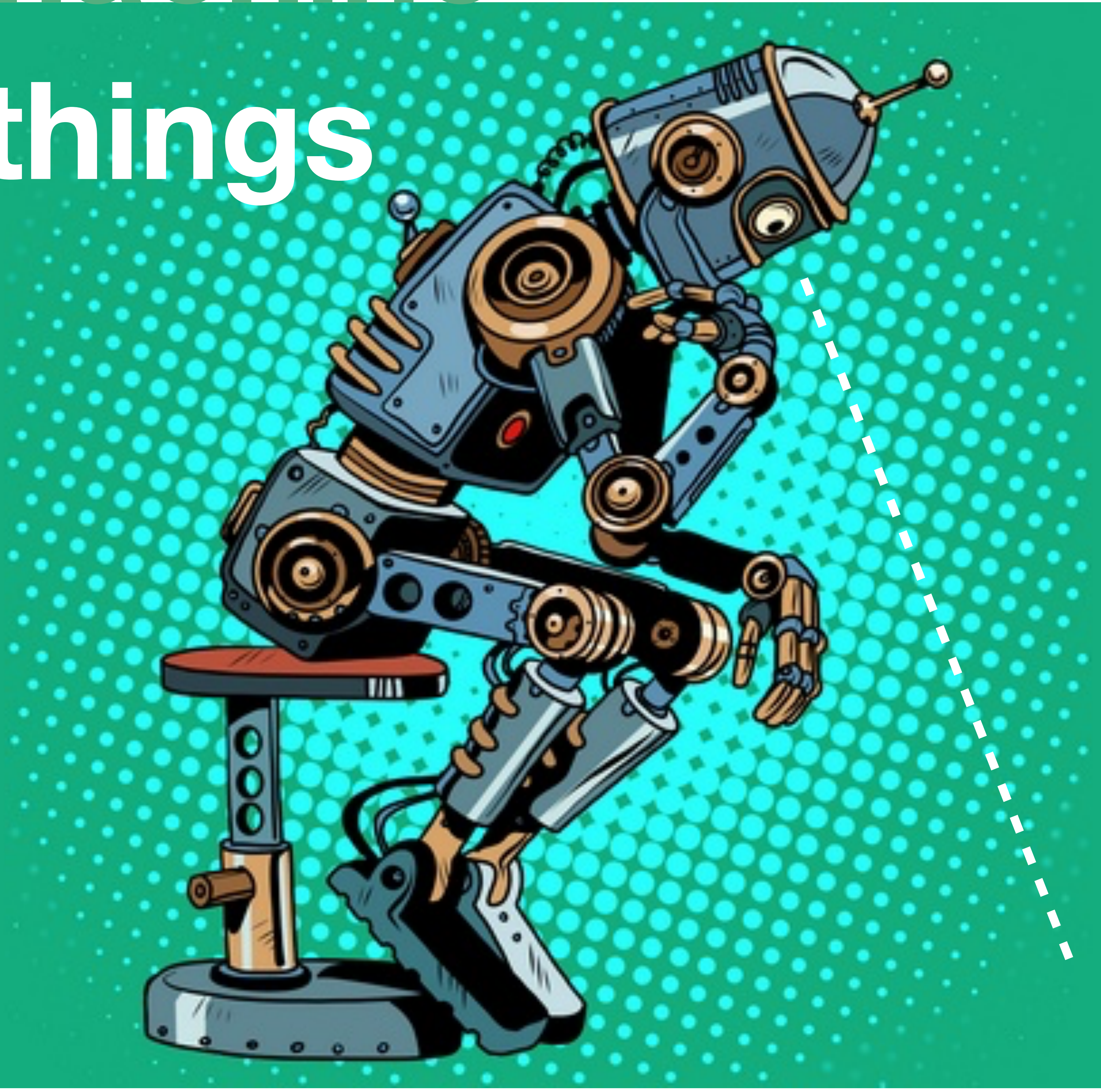
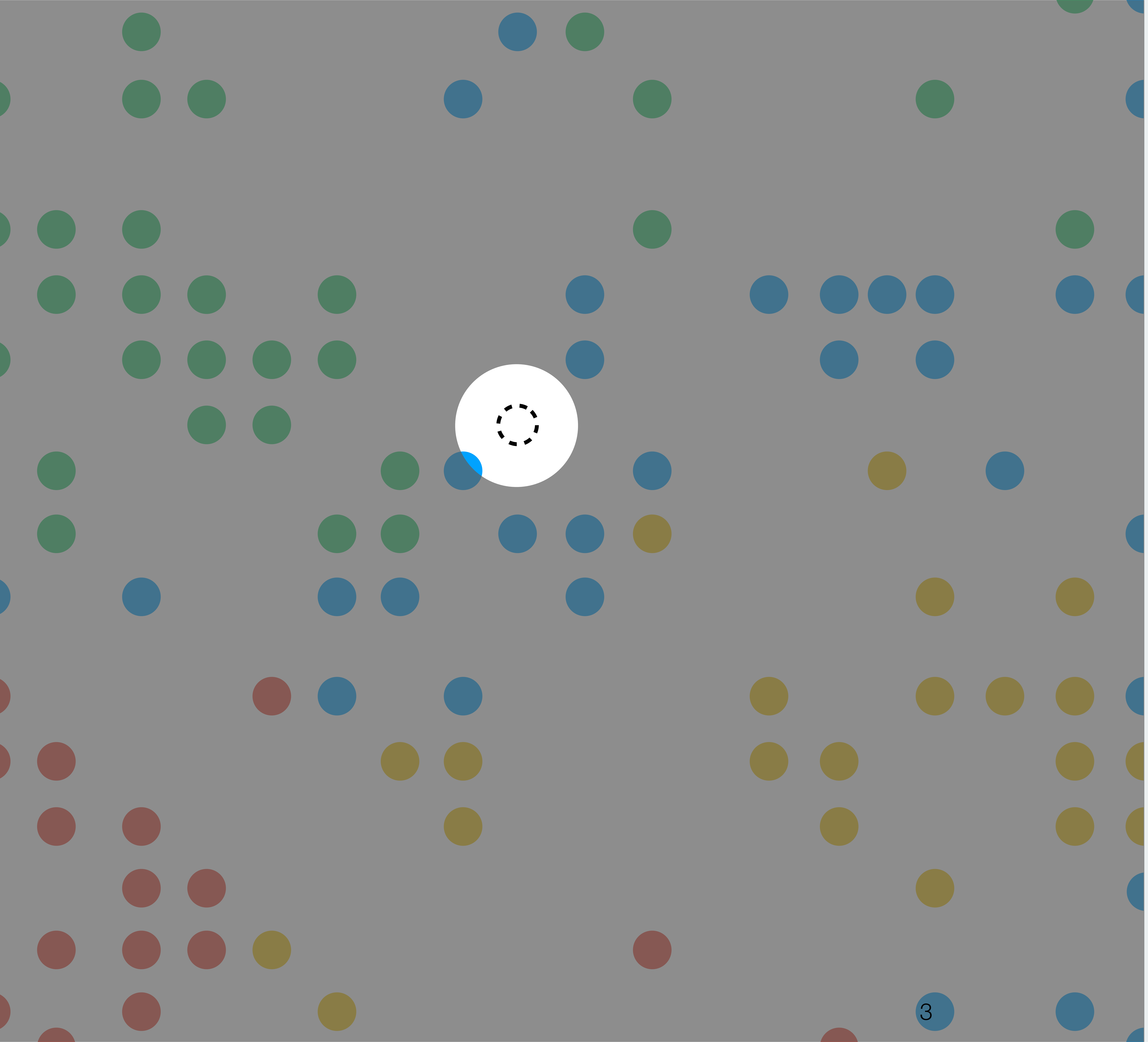
How does the machine

Learn the new things from data?



C.S Wang
2019. 07. 18

K Nearest Neighbor

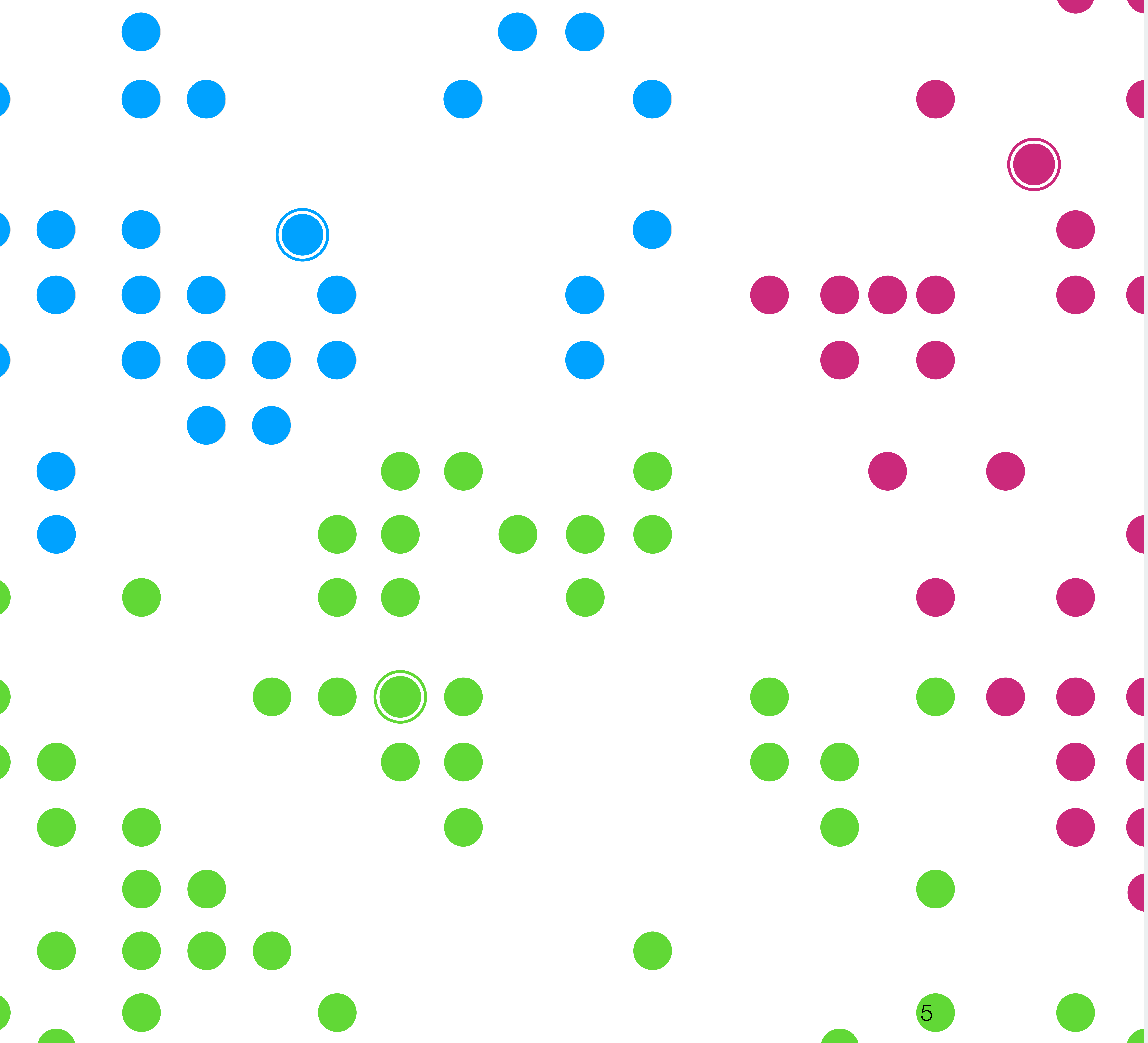


K Nearest Neighbor

- K is the **Parameter**
- Depend on Top-K neighbors
  Ensemble
- Distance Calculation

  ex: Euclidean distance

K-Means



K Means

without Label

- K is number of clusters
- Clustering the data together
- Distance Calculation

ex: Euclidean distance

Naive Bayes

Likelihood Prior

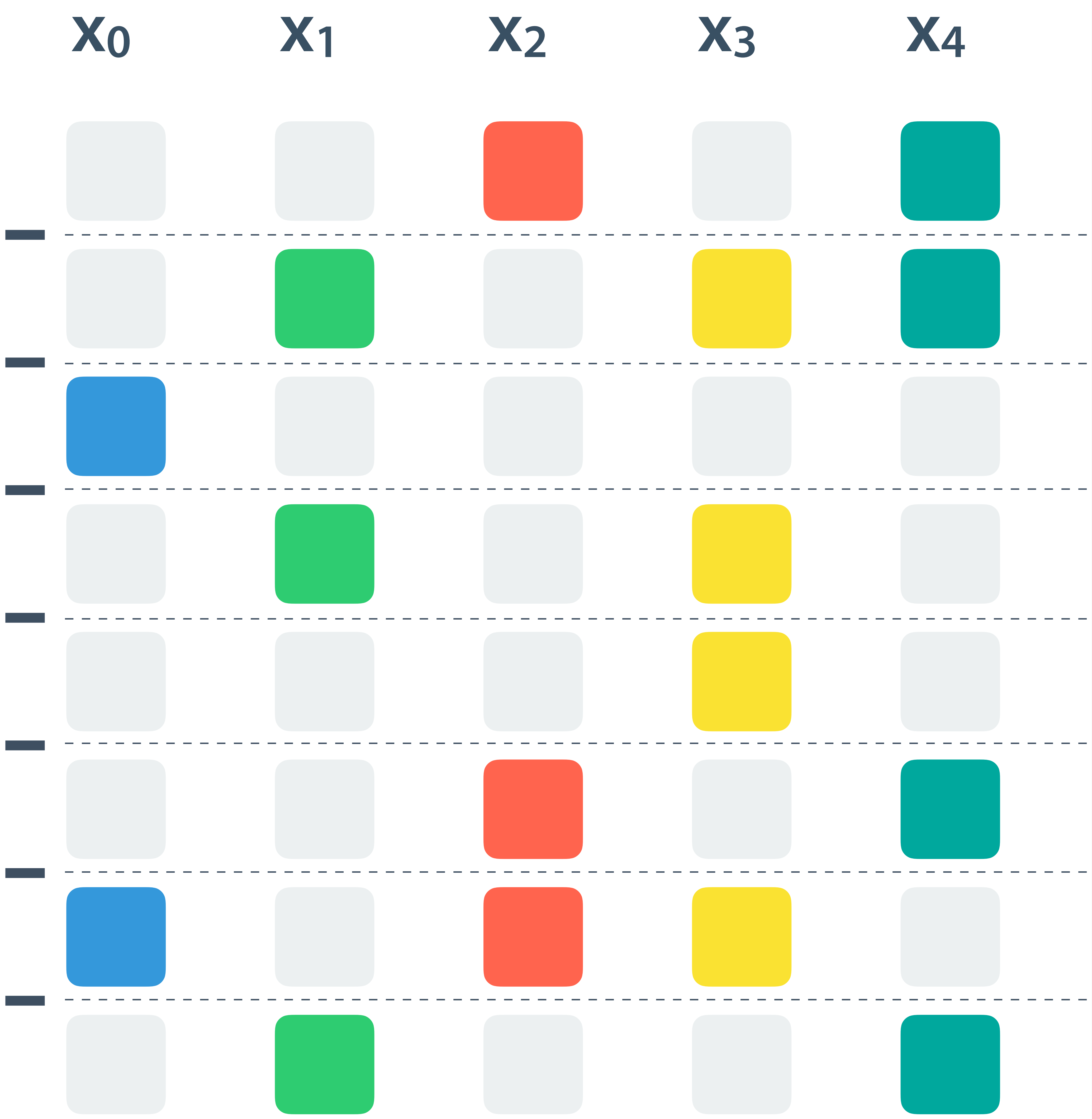
Posterior

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Evidence

Before we start…

Some things you should know. One of most important thing is **creating a table** for your data.



Basic things

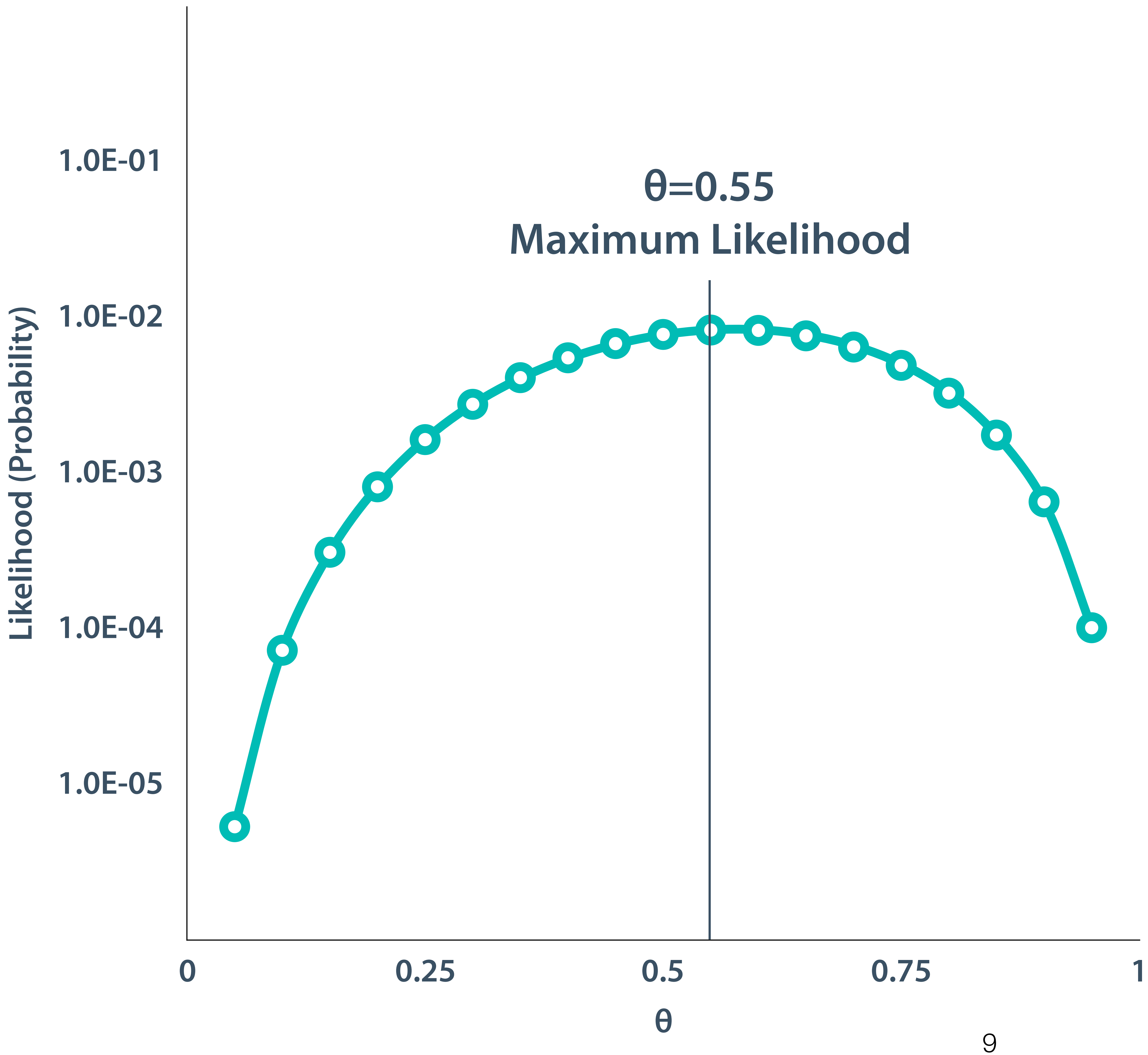
Step by Step

Prior

+	-
390/1000	610/1000

Likelihood

+	$\frac{50}{150}$	$\frac{70}{230}$	$\frac{250}{500}$	$\frac{190}{191}$	$\frac{250}{660}$	$\frac{270}{530}$
-	$\frac{100}{150}$	$\frac{160}{230}$	$\frac{250}{500}$	$\frac{1}{191}$	$\frac{410}{660}$	$\frac{260}{530}$



Likelihood

Informal, as a synonym for **probability**

- **How well** the data summarizes these **parameter θ**
- It represented by L(θ|x)

$$L(\theta\,|\,x) = P(x\,|\,\theta)$$

- Toy Example

Observed Data: HHTHTTH

What is θ ?
How to estimated it?

+(Positive) or -(Negative)?

Naive Bayes

Inference

$$P(y\,|\,x) = \frac{P(x\,|\,y)P(y)}{P(x)}$$

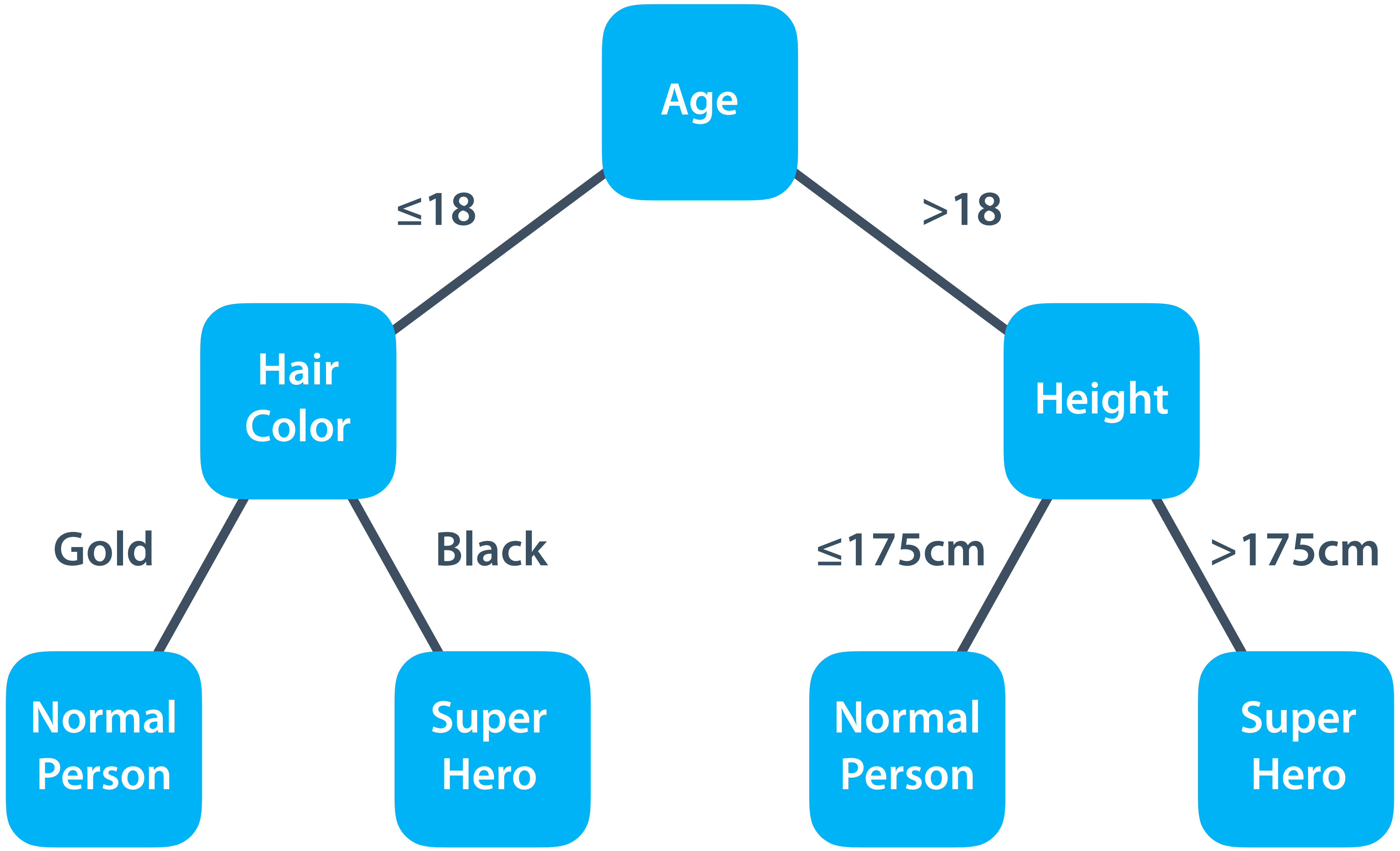
Prior

+	-
390/1000	610/1000

Likelihood

+	$\frac{50}{150}$	$\frac{70}{230}$	$\frac{250}{500}$	$\frac{190}{191}$	$\frac{250}{660}$	$\frac{270}{530}$
-	$\frac{100}{150}$	$\frac{160}{230}$	$\frac{250}{500}$	$\frac{1}{191}$	$\frac{410}{660}$	$\frac{260}{530}$

Decision Tree



Decision Tree

To see how to make a decision

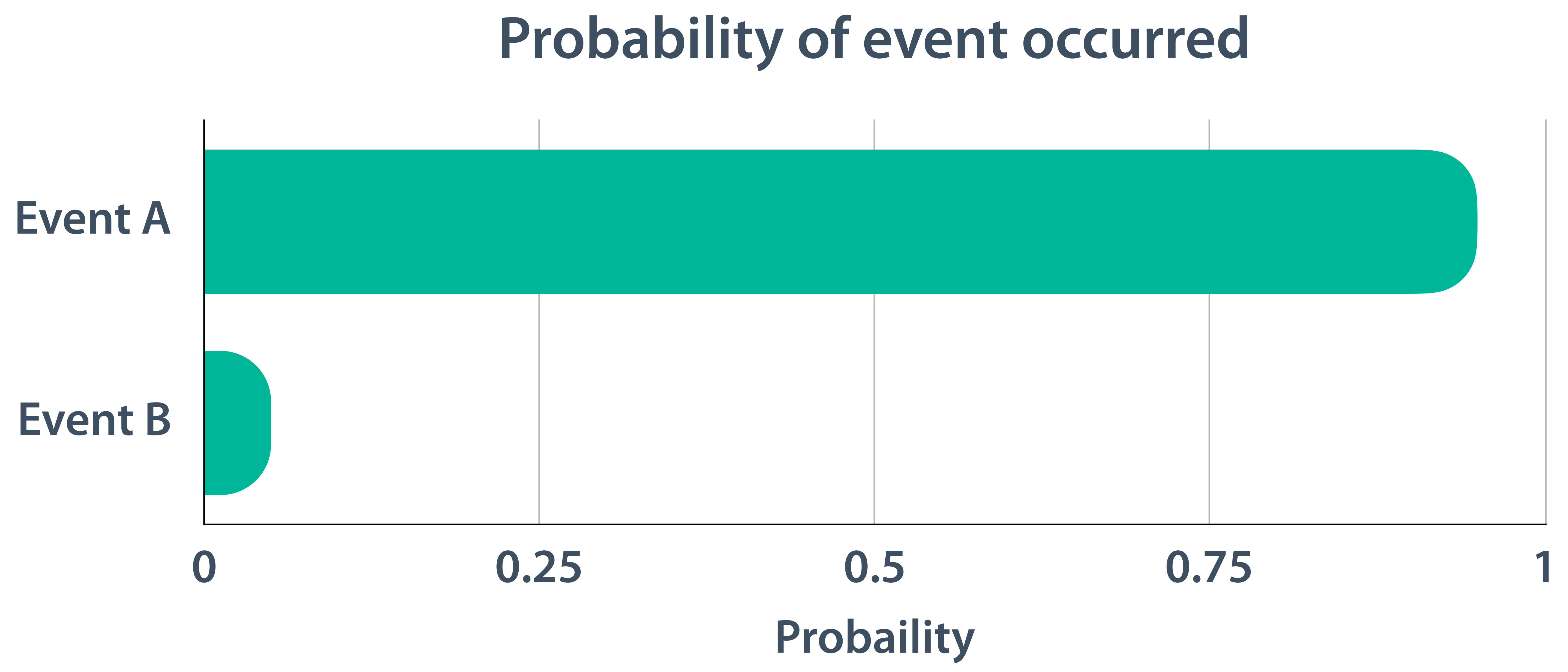
- **First, the data should be discrete**

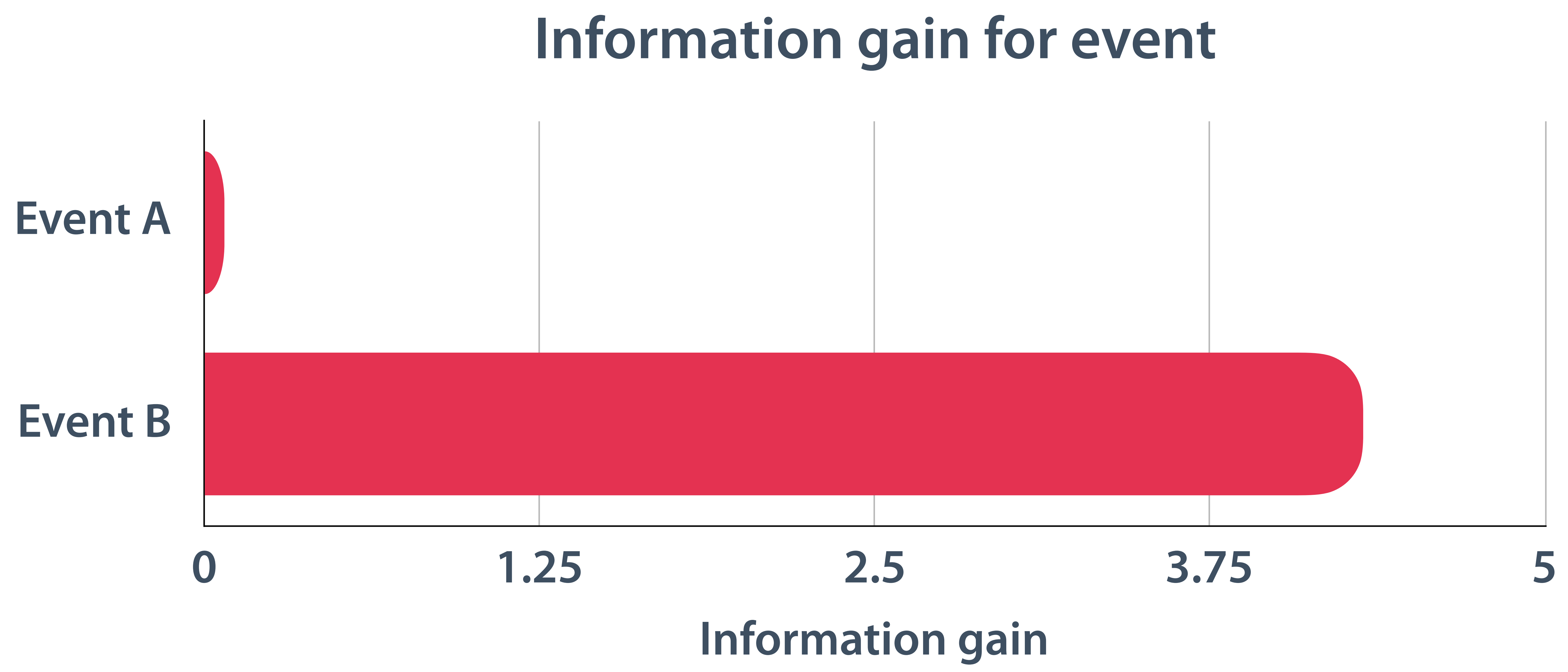
  Transforming it, if it was continuous

- **How to find the suitable node to split it?**

  Information gain & Entropy

- **Basic idea! Information**





Information Gain

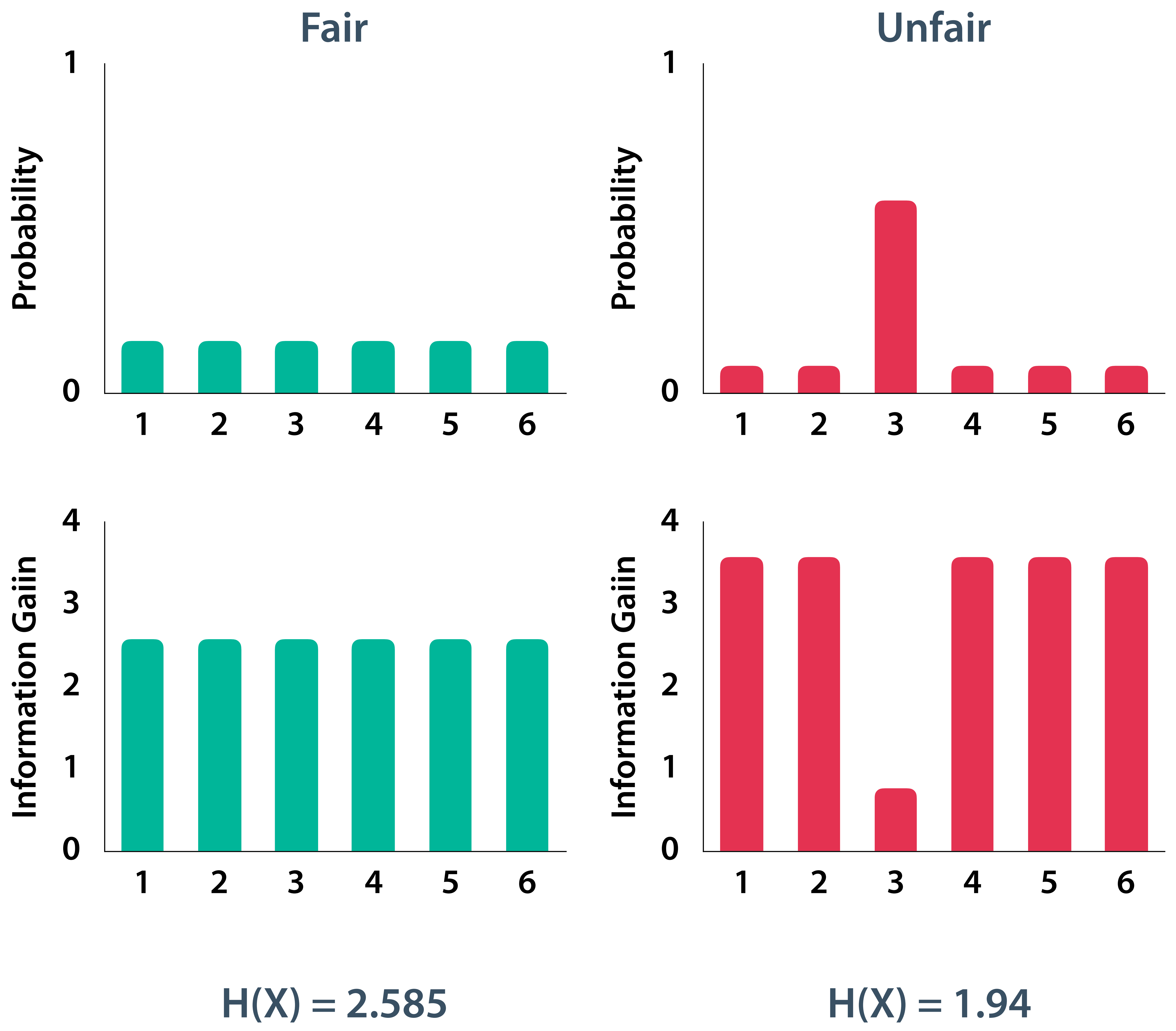
Evaluated your information

- **Which is more important?**

  Event A has 95% to occur
  Event B has 5% to occur

- **(Shannon) Information Gain**

$$IG(E) = -\log P(E)$$



Entropy

Information overall uncertainty

- Expectation of information gain

$$H(X) = -\sum_{i=1}^{n} P(x_i) log P(x_i)$$

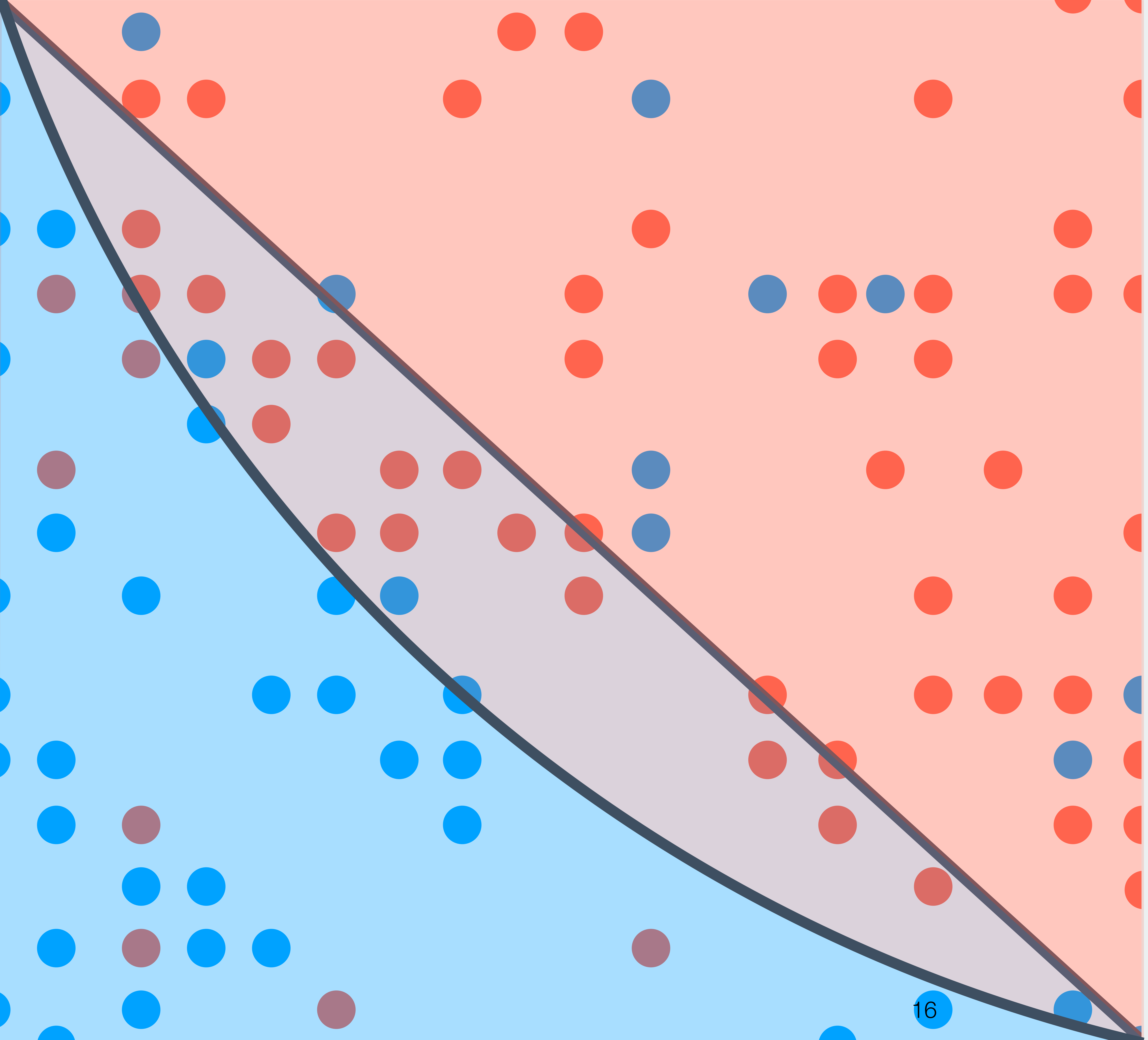
- Fair dice vs unfair dice

	1	2	3	4	5	6
Fair	1/6	1/6	1/6	1/6	1/6	1/6
Unfair	1/12	1/12	7/12	1/12	1/12	1/12

- Does large value means better?

Noooooo!! it's more uncertainty

Logistic Regression



Logistic Regression

The significance of weight

- How to split these data into two parts?

  Assume this is a binary classifier

- Logistic Function

$$h(x) = \frac{1}{1 + e^{-W^T X}}$$

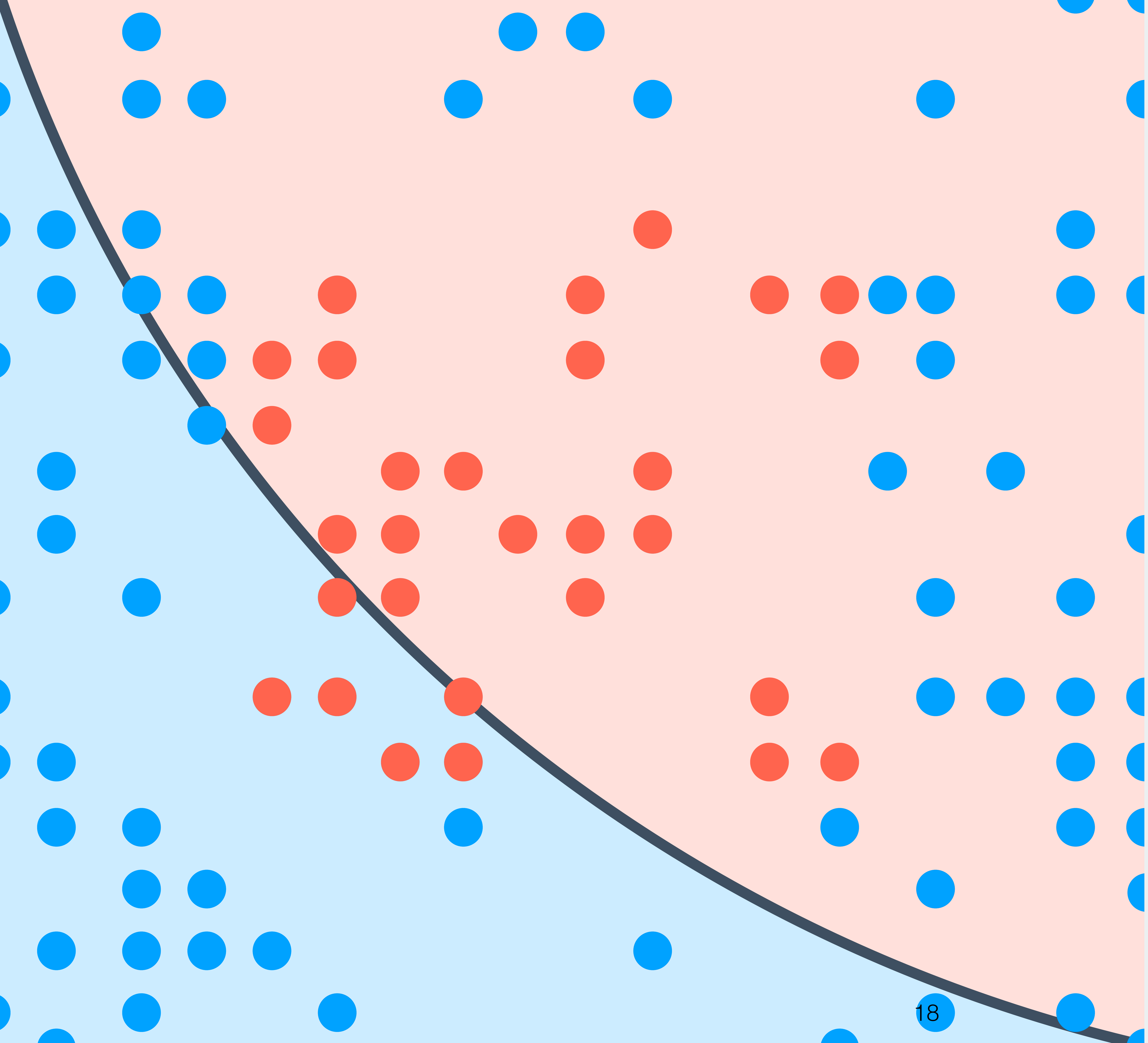
- Weight update

  Gradient descent

- How about linear regression?

$$h(x) = W^T X + b$$

Support Vector Machine



Support Vector Machine

Going into high dimension space

- Why we need this?
  (Don't forget logistic regression)

- Support Vector & Support Plane

- How to calculate the similarity in high dimension space?

  Using **Kernel Function**
  ex: linear, Polynomial, RBF